DOCUMENT RESUME

ED 202 888 TM 810 311

AUTHOR Keech, Catharine Lucas
TITLE Neglected Areas in Evaluating Writing Performance.
PUB DATE 17 Apr 81
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Evaluation Criteria; *Program Evaluation; Testing Problems; *Writing (Composition); *Writing Evaluation; *Writing Skills

ABSTRACT

The heavy concentration of time and funds to measure writing performance is the major reason other areas deserving scrutiny are so often neglected by evaluators. Three failings typical of the field of writing assessment as it is conducted for the purpose of program evaluation are: (1) a failure to view writing as a multiple construct; (2) a failure to treat writing as a process; and (3) a failure to get the most or the best information from writing samples. Each of these failings is briefly discussed. Among other things, a good writing program may change the way a student uses his writing time, helping the student begin sooner, devote more time "in toto", apply invention strategies, revise at a deep rather than merely a surface level, and proofread with greater awareness of the audience's needs. The natural irregularity of performance during development as new skills are being acquired is familiar to cognitive psychologists, but evaluators of writing programs have not typically learned how to take into account the fact that a student may score high on specific criteria during the pre-test, and score low on the same criteria during the post-test. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED202888

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper prepared for AERA, April 17, 1981, Los Angeles
Session Number 43.26:
Issues in the Evaluation of Composition Instruction

NEGLECTED AREAS IN EVALUATING WRITING PERFORMANCE

Catharine Lucas Keech
Bay Area Writing Project
University of California
Berkeley, CA 94720

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. L. Keech
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 311

Catharine Keech
Bay Area Writing Project
University of California
Berkeley, CA 94720

NEGLECTED AREAS IN EVALUATING WRITING PERFORMANCE

A Bay Area Writing Project teacher I work with commented recently, when she heard me agonizing about how to evaluate a writing program to be funded by Title IV-C, "Well surely you can show them how much good the program did -- just look at the kids' writing." It remains true that whatever else evaluators look at, in the last analysis what counts is what happens to kids' writing. The heavy concentration of time and funds to measure writing performance is the major reason other areas deserving scrutiny are so often neglected by evaluators, as my colleagues here have suggested. My teacher friend expressed the view of many evaluators of writing programs when she said, "Just look at the kids' writing." But, as I tried to explain to her some of the issues and dilemmas hidden in that casual word "just," I began to believe that many of the problems we have encountered in trying to measure program effects and outcomes have been a result of our failure in some cases to "just look at the kids' writing."

If there is time, I would like to address the issue of why it is so important to look at kids' writing, directly, rather than relying on indirect measures of writing performance, whatever psychometricians may tell us about the superior reliability or predictive validity or the relative costs of so-called "objective" measures of writing ability; but since that argument may seem like beating a dead horse to many of those here, I will first address the problem of how we look at kids' writing, and discuss some areas we have neglected in our direct measures of writing performance. "Just looking at kids' writing" has not typically been done very thoroughly or with much imagination or insight by

evaluators. What we have really done has been testing kids' writing, or scoring kids' writing, but not really looking at kids' writing -- either the products or the processes -- with a view to gaining information about what happens as a result of a writing program. We have asked the question, "How do kids' scores improve on given measurement instruments?" rather than the question, "Has this program affected student performance in observable ways?" or the question, "How have students changed their writing performance during the course of this program?" The limits of the measuring instruments typically used cause evaluators to neglect important kinds of information, and increase not the alpha error -- the likelihood of finding positive effects where in fact there are none -- but, unfortunately for the cause of writing instruction, have increased the likelihood of beta errors -- the chance that our instruments will show no positive effects of a program, although in fact important changes have occurred.

Reviewing the writing assessments conducted by many states or districts, or even by professional evaluators of special programs, I find three important areas of neglect that explain in large measure why these assessments do not answer these deeper questions, often producing inconclusive or confusing results that condemn potentially good writing programs to extinction.

Although there are notable exceptions to each of the following charges, I will assert three failings I believe are typical of the field of writing assessment as it is conducted for the purpose of program evaluation:

1. We have failed to view writing as a multiple construct.
2. We have failed to treat writing as a process.
3. We have not gotten the most or the best information from our writing samples.

Let me elaborate briefly on each of these in turn.

1. We have failed to view writing as a multiple construct.

One of the charges laid against the ETS model of holistic scoring by Cooper and Odell is that this scoring system, which is widely used in program evaluation, assumes that there is essentially only one kind of writing ability -- or that one kind of writing performance can be measured to adequately predict writing ability in all areas.[1] This view is sometimes expressed as, "Good writing is good writing. Kids can do it or they can't." I think with the advent of recent work in discourse theory and research on the very distinct kinds of structures characteristic of different kinds of writing, professional testers and evaluators have come to agree that there

---

[1] Charles Cooper and Lee Odell presented a comparative critique of approaches to writing assessment in a paper prepared for AERA in 1978. Their argument that holistic scoring is based on the premise that there is only one kind of writing is, I believe, misleading. It is as a correction to this "failing" of holistic scoring that Cooper, Odell, and others advocate the move to Primary Trait Scoring (PTS) now in use by the National Assessment of Educational Progress, which builds on current discourse theory to determine the characteristics or primary traits of pieces of writing which have different discourse functions, for instance, argument and persuasion, autobiography, self-expression, information, complaint, etc. While it may be entirely desirable to score papers for specific traits, and it is certainly necessary to develop very specific kinds of tasks to elicit narrowly defined kinds of discourse if one wishes to score the tasks in this way, the converse requirement does not hold: one may develop tasks aimed at eliciting writing for specifically defined occasions, but may score these written products holistically, or naturalistically, training readers to assign a single rapid impression marking given their response to how well students accomplished the particular task they had attempted. Cooper himself (SUNY, 1979) uses the holistic score as a touchstone to record over-all effectiveness of a piece of writing and to allow selection of samples of successful and unsuccessful writers for further analysis, although the writing task is not at all typical of those used in earlier holistic assessments. PTS and holistic scoring are both applicable to any test written in response to a clearly defined writing task. A better argument for PTS or other analytic scoring systems is offered below, in my discussion of how to get more information from our writing samples.

are several different kinds of writing -- all of which require very different kinds of writing strategies or cognitive and communicative competencies. While ETS may argue that, for purposes of predicting performance in college, it is only necessary to sample from one representative domain, I think it would be false to accuse them of arguing that other domains don't exist.

As writing program evaluators, our purposes are generally both larger and more precise than making rough predictions about which students might do well in college. We are asking, remember, very specific questions about how and what students have learned as a result of practice and instruction in a particular program. The important implications of recent discourse theory are not that we should collect samplings of student performance in every conceivable kind of writing situation. Nor are they merely that we need "at least two" samples of writing from each student.[2] Rather, the implication is that it matters what kind of task we create on our writing test if we are to record actual program effects. Imagine a writing program that has concentrated all year on teaching students the rudiments of expository writing, urging development of an introductory thesis and the use of explicit generalities backed up by details, examples, or arguments. The teacher has has worked hard with students on the internal coherence of paragraphs, and on the use of such transitional expressions as, "however," "therefore," "on

[2]In the ETS, Diederich model, more than one sample of writing was recommended for testing, but the purpose in this case was to increase the reliability of the student's score. To increase reliability, the second sample should be parallel in task demands and discourse type to the first sample. If we instead draw two samples of writing, one from each of two different discourse types, for example, a narrative based on personal experience, and an argument based on a current issue, as is done by the ETS in the California High School Proficiency Exam, what is gained, instead of increased score reliability, is a broader and more representative picture of the student as a writer.

the other hand," "in the same way," and "for example," to guide the reader through the labors of an informative or argumentative piece of prose. Imagine now that a program evaluator has created a pre-post test writing assessment design, using the following topics:

1. Describe something you enjoyed doing as a child.
2. Describe an experience from which you learned a lesson.

Nothing in the teaching program has helped students to improve the particular skills required to produce better narratives. In fact, many of the skills being worked on, if applied to the narrative task, would be likely to cause the student to produce a less successful paper . I have seen it happen. In the tenth grade, a student's holistic score on a traditional personal-experience task dropped strikingly from his ninth grade performance, because the second time around he approached the writing test with his newfound expository skills, producing a tightly organized essay which totally lacked the vivid details or lively personal voice of his loosely constructed ninth grade piece. Looking only at his scores, an evaluator might conclude that the tenth grade program not only failed to teach this student to write better, but actively hurt his writing ability.

The moral of the story is simple: construct writing tasks that reflect the aims of the program. In earlier grades, when the aims of writing instruction may be to help students acquire basic scribal automaticity, the ability simply to transcribe their thoughts onto paper, a writing test might be one of general fluency, designed, like traditional holistic assessments, to elicit lively writing and as much writing as possible. The appeal to personal experience on such tests is a way of assuring that the student will have access to ample material out of which to construct a text. This does not mean that

first or second grade is too early for teachers to begin to elicit both talking and writing that fulfills a variety of functions in addition to self-expression or personal narrative. Research is needed in classrooms where students begin at early ages to write about information and ideas, to determine the extent to which their composing strategies and language choices differ for these tasks, compared to traditional English theme tasks, like, "describe your favorite object." It does mean that general self-expression tasks may be quite adequate to measure basic fluency and mastery of surface conventions of written English as well as measuring one important function of written language, but they are not adequate to measure other kinds of discourse planning skills or the full range of semantic and syntactic resources older students will need to accomplish other important functions of written language.

To summarize: we need not simply more samples of writing on our writing tests, but more carefully designed sampling of the kinds of writing actually addressed by the writing program being evaluated.

2. We have failed to view writing as a process.

Just as we need to recognize more than one kind of written product which might be examined as a measure of program effects, so there is more than one way students may produce the written products we evaluate. One of the most important outcomes of a good writing program may be to change a writer's processes, his approach to a writing problem, in ways that will bring about differences in what he is able to achieve. Among other things, a good writing program may change the way a student uses his writing time, helping him begin sooner, devote more time *in toto*, apply invention strategies, revise at a deep rather than merely a surface level, and proofread with greater awareness of his audience's needs. Now, suppose a writing program has accomplished these

remarkable and desireable things. But the program evaluator, because he is using established procedures, samples only one kind of writing process: the process that occurs when a studentsits down to an unfamiliar topic, reads that he has thirty minutes to write, dashes off what he can, and turns it in without re-reading because that is what he thinks must be expected when he is given so little time, compared to the time he is offered in which to complete class assignments. Before this particular writing course, he, like hundreds of other school children, used this very same process no matter what the task, and no matter how much time was allowed. In fact, if he had been given extra time earlier on to write an essay on this topic, he might have written worse, since he would have thought he was supposed to just keep scribbling from the start gun to the finish line. Since his writing course, however, he approaches his writing differently. The thirty minute stricture on a topic which he sees has interesting possibilities is confusing. Suppose he mistakenly tries to apply his new process to this test occasion. In one class where I gave a 30-minute test to students who had begun to make such changes in their processes, the majority of the class sat silently, making a few notes or staring out the window for anywhere from ten to twenty minutes, before beginning to write. What they wrote in the final ten minutes was good--as good as what they wrote without much planning early in the year, and as good as what is generally gotten from students on tests like this one. What it was not (and this is critical for measuring program effects): it was not even close to as good as writing I pulled from their class folders, writing on similar topics but with much longer time allowed.

Please notice that I am not saying that all tests of writing ability must provide unlimited time for writing. Nor am I saying that students will always do better when given more time. I have observed cases where students wrote worse because they didn't know how to use additional time well. What I am

saying is that when a writing course changes how well students are able to use writing time, we will not be able to measure or observe that change if we give only short impromptu tests. As before, if our conventional measures don't allow us to show what has really happened to a student's writing process, we may well reach the erroneous conclusion that the writing course has actually hurt, or at best failed to help, students' writing performance. The thirty-minute post-test in the class I observed showed some students doing less well than they had on such tests at the beginning of the year: because they cared more about their writing, because they had learned how to define writing tasks more complexly and richly, because they wanted to attempt more, they spent too long planning and were rushed when they wrote. They had temporarily lost touch with the proper degree of casualness required for successful shoot-from-the-hip theme-ing.The writing teacher could fairly argue that in her view their mediocre performance on this test represented a step forward for these students. Putting ceiling on the time students spend on test items and the resources they have access to in composing for test purposes puts definite ceilings on the quality of work they may produce. For the traditional testing purposes of agencies like ETS, these limits are useful. They cause student performance to be spread across a continuum which allows us to identify which students do the task best and fastest. The limits also produce short, easily scored pieces of writing, which are perfectly adequate as a rough estimate of a students' ability to perform in writing in the future.

But look again at the question program evaluators are asking. They need to know not which students do the task best and fastest, but how many students can do the task at all compared to performance by other groups or by these same students at the beginning of the year. They also need to know <u>how well</u> can the very best writers do the task, and <u>how well</u> can

the worst writers do the task, given the strategies and resources students have gained from the writing program. Has the program changed the number of students who can do the task, or the degree of successfulness of students at different ability levels?

In order to answer these questions, evaluators must sample writing done in several settings. It may be an important instructional or evaluational goal that students improve performance on short timed tests: if so, such writing tasks should be included. But if this is not the sole goal of the program--and I submit that few writing programs are doing their job if their goals are limited to improving scores on such tests--then writing samples should be drawn which reflect the student at work composing in more naturalistic settings.

This can be done in several ways. Most obviously, samples of classroom work, portfolios of student writing across a year, can be analyzed to show differences in how students defined and attempted to solve different writing tasks. But purely naturalistic samples make problems for evaluators who wish to draw comparisons between classrooms and schools, or from the beginning to the end of a year. Alternatively, experimental test situations may be designed to simulate more naturalistic writing processes, allowing students to demonstrate the fullest and richest processes aimed for in the writing program. A district might provide in its assessment plan a model for teaching the composing process by implementing a four-day test sequence; allowing for prewriting exercises, including discussions and note-taking, on <u>day one;</u> drafting and composing individually, on <u>day two</u>, with reference only to notes taken the day before and left in class; sharing drafts and collecting feedback responses from peers on <u>day three</u> (again leaving all drafts at school); and finally, on <u>day four</u>,making final revisions and turning in the work.

Such a test-design, administered at the beginning of the year, would provide excellent diagnostic information at several levels: teachers might examine student work on each of the four test days and be able to form an impression not only of how students use peer feedback, but also of the quality of help they are able to offer others. Repeated at the end of the year, a similar test occasion and examination of notes and drafts could provide evaluators with evidence of change in students' richness of problem formulation during pre-planning, their incorporation of pre-planning in initial composing, and their ability to give and use evaluative responses constructively in revision.

Versions of this test design are being tested by BAWP teachers and schools, but evaluators of large-scale programs might quail at attempting it widely. Imposed from outside the school, the plan could cause resentment among teachers who might see the extra days of testing as time taken away from instruction--unless they were already trained to perceive this as an important kind of writing instruction in itself. And of course the costs of analyzing not merely a short piece of writing, but a whole set of drafts written by several hundred or several thousand students would be astronomical. There is an obvious solution to the cost problem: careful sampling of selected student work for close analysis, but that begins to encroach on my next point, how to get better information from our writing samples.

Before moving on, let me just summarize my point about writing processes: look at the changes in writing process which might be occuring as an intended result or as a side-effect of the instructional program you are evaluating. Then look at the writing test you intend to use. Is there any chance that changes in writing process will actually be reflected in improved performance on the test you are considering? Most test procedures currently in use would have to be ruled out on this grounds alone.

3. We are not getting the most or the best information from our writing samples. Having collected actual samples of student writing--either in experimental test situations or from classroom portfolios or assignments--what information can be gained about student abilities and the effects of the writing program?

Typically, evaluators have simply scored writing samples with a single rating for rhetorical quality, as in so-called holistic scoring, or with a set of ratings describing the presence or absence of certain traits, as in primary trait scoring, or have used analytic scales like the one developed by Paul Diederich and used at the Center for the Study of Evaluation, as a way of breaking out particular rhetorical features which contribute to the overall rhetorical success of a piece of writing. A review of the relative costs, reliability, validity, and kinds of information gained from these various assessment methods appears in the publication, Direct Measure of Writing Skill (CAPT, 1980).

Each of these measures has important limitations, but all can be useful, especially when used as one of several ways to look at student writing. Users of all of these measures report increasing reliability of scoring procedures as writing task specifications and reader training become more refined and precise. But all of these measures are subject to the natural and often extreme variation in student performance on different writing occasions which occurs as a result of several factors beyond the test-makers' control and which produces scores on writing tests which can be baffling and frustrating to students, teachers, and evaluators who believe a program has actually helped students improve their writing. Something is clearly missing in the information we are getting from these measures--something that might help us interpret why developing writers appear to perform so inconsistently over time.

An obvious criticism of simple summative measures like holistic scores is that a single score for effectiveness, while providing useful information about how a student has performed generally in comparison to his peers, provides little information about the student's actual strengths and weaknesses as a writer. In a laudatory attempt to get more information about particular characteristics of student writing, some evaluators add error counts or rating for mechanics to the holistic score. I would like to warn against any purely quantitative tallying of errors of surface convention as being not only unlikely to yield useful or relevant information about student writers, but as being actively misleading or harmful if unaccompanied by some qualitative analysis of those errors. Again and again, correlational studies have shown that many kinds of errors play virtually no part in determining a reader's response to a piece of student writing, although general qualitative ratings on mechanics do explain some of the over-all effectiveness of a piece. What the qualitative ratings do that simple error counts don't do is determine whether the errors are of the kind that really interfere with readability, whether they are of the kind that signal a low or a high level of difficulty in what the writer is attempting, and whether they appear to mark patterned disfunctions in the writer's production of written language or appear, given the presence of well-formed words and sentences, to be mere slips of the pen from hasty transcription. To evaluate student performance meaningfully and to understand positive program effects, we must do more than count errors--we must figure out what errors mean.

Further, the analysis of errors must be accompanied by other kinds of analysis if we are to explain why large numbers of students so often

fail to improve their performance in neat, stepwise score increases on our current measure. The natural irregularity of performance during development as new skills are being acquired is familiar to cognitive psychologists. Piaget describes learning as occuring in alternating periods of equilibrium and disequilibrium. But evaluators of writing programs have not typically learned how to take into account the fact that a pre-test measure, taken during a period of equilibrium, when the student is easily able to do all that he attmpts, may score high on such criteria as correctness, flow, voice, diction, or other rhetorical features, while a post-test measure, taken when a student is attempting to apply skills not yet fully assimilated or to accomplish tasks more difficult than any he has attempted before, may score low, showing a fresh rash of errors, inadequacies and awkwardnesses not apparent in the earlier writing. The fact is that some occurences of errors or rhetorical flaws are actually signs of a writer's progress in developing greater communicative competence.

There may be several different phenomena involved here. It is possible that in developing writers, attention to new writing strategies, for instance concern to produce more varied sentences, or concern to organize a piece of writing with a clear beginning, middle, and end, simply requires so much attention that the student temporarily abandons close attention to details which had previously concerned him, for instance spelling or capitalization. An increase in surface errors, so often regarded by teachers as a sign of character deterioration in once meticulous students, may be perceived instead as an inevitable accompaniment to a new level of attempt, a shift in focus of attention, a temporary lapse in applying editing skills which will be remedied when the new composing skills become more automatic. Alternatively, the new skills may interfere

with old knowledge and require some time to sort out, as Maimon and Nodine demonstrate in their study of students whose training in sentence combining resulted in increased syntactic complexity, but also in higher incidence of run-ons and sentence fragments not apparent before training. (1978)

There is another way we might regard this phenomenon of errors which signal advances for a writer. There are some errors a writer is simply not going to make until he reaches a certain level of advancement: syntactic errors, like dangling modifiers can't appear until the student begins to produce sentences containing participial phrases; punctuation errors like a failure to mark non-restrictive clauses with commas can't appear in the papers of students who don't write relative clauses at all; spelling errors that occur in multisyllabic words can't appear in the writing of mono- or di-syllabic students. The absence of error common in papers of conscientious students whose writing is restricted to simple syntax and vocabulary cannot be regarded as signalling a high level of writing competency.

An important study by Freeman and Pringle called "Indices of Growth in the College Years" (1980) shows that students, as they go through college learning the languages of their various disciplines and tackling increasingly difficult cognitive tasks in their written work, appear to make no progress when their writing is judged for its rhetorical effectiveness, beauty or style. But these researchers developed another measure to record what the student was attempting in terms of how much information was being synthesized at what level of abstract thought. On this measure, it appeared that students were improving markedly in an important set of composing skills: the ability to deal with increasingly complex ideas and to synthesize greater amounts and more kinds of information in ways appropriate

to particular disciplines. If we regard this kind of cognitive development as a component of communicative competence--and I think we must--we must also realize that as evaluators we need to analyze student texts for signs of development judged by cognitive as well as rhetorical criteria.

Clearly then, evaluators must join researchers in examining student texts at the micro-level, working with teachers to build a new understanding of what features of writing commonly regarded as rhetorical flaws may be associated with advances to higher levels of communicative intent and attempt. Although we cannot take the approach that any increase in flaws is to be regarded as a sign of growth, I submit that we must learn to recognize some occurences of new flaws as heralding growth just as clearly as does a decrease in error or removal of old flaws.

Practically speaking, although it is important to move toward micro-analysis of texts and toward developing new and better ways to interpret patterns of error and determine evidences of development in student writing performance, the costs of such analysis are fobidding, especially if we remain wed to the old evaluation-cum-research models that require hundreds of subjects to be tested with identical measures which yield information easily reducible by summary statistics. A better model for assessment is presented by Cooper (1979) in his report on the writing abilities of entering freshmen at SUNY, Buffalo. In addition to descriptive statistics of student performance on global measures, he draws carefully selected samples for close analysis of discourse features. This approach allows evaluators to search for real program effects on the writing of individual students, providing important information which allows them to interpret the global findings

With any luck, evaluators may soon abancon the old battle over direct versus indirect assessment and enter the new battles, fighting for outcome measures which allow us to really "look at kids' writing": viewing writing as a multiple construct; viewing writing as a process; and analyzing texts to provide more and better information about changes in student performance as developing writers progress along the bumpy, circuitous road to full communicative competence.